Cross-ethnic measurement equivalence of measures of depression, social anxiety and worry.

By: James P. Hambrick, Thomas L. Rodenbaugh, Steve Balsis, Carol M. Woods, Julia L. Mendez and Richard G. Heimberg

Hambrick, J.P., Rodebaugh, T.L., Balsis, S., Woods, C.M., Mendez, J.L. & Heimberg, R.G. (2010). Cross-ethnic measurement equivalence of measures of depression, social anxiety and worry. *Assessment, 17*(2), 155-171. doi: 10.1177/1073191109350158

Abstract:

Although study of clinical phenomena in individuals from different ethnic backgrounds has improved over the years, African American and Asian American individuals continue to be underrepresented in research samples. Without adequate psychometric data about how questionnaires perform in individuals from different ethnic samples, findings from both within and across groups are arguably uninterpretable. Analyses based on item response theory (IRT) allow us to make fine-grained comparisons of the ways individuals from different ethnic groups respond to clinical measures. This study compared response patterns of African American and Asian American undergraduates to White undergraduates on measures of depression, social anxiety, and worry. On the Beck Depression Inventory—II, response patterns for African American participants were roughly equivalent to the response patterns of White participants. On measures of worry and social anxiety, there were substantial differences, suggesting that the use of these measures in African American and Asian American populations may lead to biased conclusions.

Keywords: cross-cultural measurement | equivalence | social anxiety | depression | worry | psychology

Article:

As a result of the underrepresentation of African American and Asian American individuals in research samples (Neal & Turner, 1991; Snowden, 2001; Wells, Klap, Koike, & Sherbourne, 1999), much of our knowledge about depression, social anxiety, and worry in African American and Asian American samples has been extrapolated from assessment methods standardized in primarily White1 samples. Such an approach can be understood in the context of the differences between emic and etic approaches (Brislin, Lonner, & Thorndike, 1973). Measurement of anxiety and depression has not typically focused on culturally bound constructs (the emic

perspective), but has instead assumed, at least provisionally, that such constructs are relatively universal (or culture-neutral: the etic perspective). The key benefit of an etic approach to constructing questionnaires is that it allows comparisons of symptom levels across groups. However, researchers may go too far in assuming neutrality. Even the most straightforward question can yield very different answers depending on what is asked; in one Latino sample, estimates of household income shifted 15% depending on whether respondents were asked how much money they sent to family members in their countries of origin (Alegria et al., 2004). Such a question is not commonly asked about household income; its formulation required specific information about a cultural context.2

Recently, the field has developed techniques that may facilitate understanding the extent to which an etic approach can be supported. Analysis of differential item functioning (DIF), when conducted within an item response theory (IRT) framework, is one such statistical technique. In this context, items from a particular measure are conceptualized as indicators of a latent variable. Figure 1 details a set of item characteristic curves (ICCs) for a particular binary item on a hypothetical measure of social anxiety. Each ICC represents the likelihood that a particular response option (e.g., “true”) is chosen at different levels of latent pathology for a particular group. If the likelihood does not differ between groups, then the ICCs lie on top of each other, and there is no evidence of measurement bias. Discrepancy between the ICCs indicates some degree of measurement bias.

Figure 1 is omitted from this formatted document.

Based on the two-parameter logistic model used to create Figure 1, analyses of DIF can define this bias according to two parameters: discrimination (or a parameter) and threshold (or b parameter). Conceptually, the discrimination parameter indicates how well a particular item relates to the underlying latent variable, and the threshold parameter indicates at what point in the range of the latent variable the two responses are equally likely. Significant discrimination DIF indicates that an item relates more strongly to the underlying construct in one group versus another, suggesting that interpreting means or effect sizes across groups may be problematic because the scale’s items are a better measure of the construct in one group than the other. The presence of significant threshold DIF, in contrast, suggests that interpreting means across groups may be difficult because differing observed means may actually indicate the same level of the latent construct and identical observed means may indicate different levels of the latent construct. Furthermore, DIF analyses control for actual mean differences that might exist between the groups, allowing for more unbiased measurement.

As a result, even full data on norms within ethnic groups (which are rare) would not necessarily allow the researcher or clinician to draw appropriate conclusions: Apparent differences could reflect measurement error in need of adjustment or true variance that should be acknowledged. Review of three commonly used measures, the Beck Depression Inventory–II (BDI-II; Beck, Steer, & Brown, 1996); the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, &

Borkovec, 1990); and the Social Interaction Anxiety Scale (SIAS; Mattick & Clarke, 1998) reveals that existing psychometric data regarding these measures does not alleviate concerns about the possibility of threshold or discrimination DIF; instead, the existing data suggest that such departures from metric equivalence are likely.

The Beck Depression Inventory–II

Throughout its various iterations, the BDI has been one of the most studied and most commonly used measures of dysphoria (Beck, Steer, Ball, & Ranieri, 1996). Studies of cross-ethnic equivalence of the various versions of the BDI have generally supported their reliability and validity, whether measured in large multiethnic samples (Carmody, 2005) or in a sample of low-income African American medical outpatients (Dutton et al., 2004; Grothe, Dutton, & Jones, 2005).

One major problem with the research on the BDI-II is that, even in studies with large samples, individuals from minority groups are underrepresented (Beck, Steer, Ball, et al., 1996; Carmody, 2005). Beck et al. wished to examine the correlation between ethnic categories and BDI-II scores but had to collapse individuals of distinct ethnic backgrounds into a non-White category because of small sample size. The resulting lack of information leads to problems of interpretation, especially considering that subsequent studies found African American individuals did not differ from White individuals on mean levels of dysphoria on the BDI-II (Carmody, 2005), whereas Asian American individuals reported significantly higher scores than White individuals (Lam, Pepper, & Ryabchenko, 2004).

Interpretation of these data is further complicated by evidence that the BDI-II might overestimate levels of dysphoria within Asian American, and at least some African American, samples. Although Asian American individuals score higher on the BDI-II than White individuals, they have also been found not to differ on other questionnaires or structured interviews measuring depression (Lam et al., 2004; Okazaki, 1997). In the studies examining BDI-II scores of African American patients in a primary medical care waiting room, 30% of respondents met criteria for at least mild depression. Although such a result is not unprecedented, the authors noted that this prevalence estimate appeared high, particularly considering the fact that the sample was not psychiatric. It seems plausible that these anomalies may be due to DIF.

The Penn State Worry Questionnaire

The PSWQ is a measure with ample evidence of its generally strong psychometric characteristics but minimal comparison of those characteristics across groups of diverse ethnic backgrounds. Three studies have examined the differences in the PSWQ scores of multiethnic samples. In two of the three studies, African American and White samples did not differ in their tendencies to engage in pathological worry (Gillis, Haaga, & Ford, 1995; Scott, Eng, & Heimberg, 2002). In the third study, African American undergraduates endorsed significantly less pathological worry

than White undergraduates (Carter et al., 2005). Scott et al. (2002) also reported that Asian American undergraduates did not differ from African American or White undergraduates in their experience of pathological worry.

These studies also looked at the reliability and validity of the measures within the groups. The internal consistency of the PSWQ within the subsamples was very high in two studies (Carter et al., 2005; Scott et al., 2002). Gillis et al. (1995) did not report internal consistency data. A factor analysis of the PSWQ found evidence for a three-factor solution in the African American sample (Carter et al., 2005), which differed from the two-factor solution found in the White sample (as well as the two-factor solution typically found in factor analyses of the PSWQ; e.g., Fresco, Heimberg, Mennin, & Turk, 2002; Hazlett-Stevens, Ullman, & Craske, 2004). Furthermore, in contrast to the responses of White participants, the responses of African American participants exhibited only mild convergent and discriminant validity with measures of depression and anxiety. Such findings may suggest that the measure's items relate to the underlying latent construct of worry differently across groups; therefore, discrimination DIF may be present.

Scott et al. (2002) also found that, on a measure of general worry, African American participants reported less worry than either Asian American or White participants. Differences on this measure, coupled with the failure to find differences on the PSWQ, might suggest the presence of threshold DIF for at least one of these measures.

Social Anxiety and the Social Interaction Anxiety Scale

Currently available evidence suggests that Asian American individuals report significantly higher mean levels of social anxiety than White participants (Hong & Woody, 2007; Horng & Coles, 2004; Lee, Okazaki, & Yoo, 2006; Okazaki, 2000; Okazaki, Liu, & Longworth, 2002). The robustness of these findings across multiple methodological approaches and multiple samples appears to support the notion of fundamental differences in the ways that Asian American and White individuals experience social anxiety. Neal and Turner (1991) reported one study indicating potentially heightened social anxiety in African American samples. However, they also found that the number of African American participants in studies of social anxiety disorder was extremely low, with a mode of zero. We found only one study of social anxiety in African American adults subsequent to the studies reviewed by Neal and Turner; African American participants achieved significantly lower scores than White participants on the Social Phobia and Anxiety Inventory, although the authors reported that this difference may have been confounded with differences in income level (Gillis et al., 1995).

For this study, we examined the SIAS because it is a frequently used, psychometrically sound measure of symptoms of anxiety in interpersonal situations. Despite findings indicating good reliability and validity of this measure in predominately White samples, as with most measures

of social anxiety, there has been little or no study of the reliability or validity of the SIAS in Asian American and African American samples. The absence of this information makes any finding of mean differences difficult to interpret; such differences could be due to differences between populations on levels of the latent construct, failures of metric equivalence, or some combination of both of these factors.

The available literature on the validity of these measures in African American and Asian American samples is difficult to interpret. On the basis of conventional approaches to cross-ethnic comparisons, such as mean differences and factor structures in separate single-group analyses, one might conclude that African American and White samples are roughly equivalent and that Asian American samples endorse greater depression and social anxiety than White samples. However, if a measure assesses anxiety differently in African American or Asian American samples than it does in White samples, reports of, for example, higher levels of social anxiety may instead only reflect biased measurement of such phenomena. Simply comparing item scores between groups does not control for true mean differences across groups. To accurately test for DIF, the likelihood that an item is endorsed by a member of one group as compared with a member of another group must be compared with the level of the latent trait held constant.

This study sought to conduct the first large-scale examination of differences in response patterns between African American and White participants, and between Asian American and White participants on the BDI-II, the PSWQ, and the SIAS. Because these questionnaires have been validated in predominately White samples, we chose to use the White sample as the comparison group for both the African American and Asian American samples.

Method

Participants

Participants were undergraduates at an urban Northeastern university and completed the measures as part of a questionnaire packet administered at the beginning of each semester, for which they were given credit in their introductory psychology class. Specific measures included in the packet varied by semester; the BDI-II, SIAS, and PSWQ were not administered every semester. In addition to the self-report questionnaires, participants described their ethnicity using one of the following identifiers: White, Hispanic, African American (or Black), Asian American/Pacific Islander, Mixed, or Other. Participants who reported their ethnicity as Mixed or Other were removed from consideration. We planned to use groups of about 300 participants or more. We selected this minimal sample size based on our experiences with similar analyses, because there is no clear guide in the literature regarding minimum numbers of participants for such analyses. In the final analyses, individuals from particular ethnic groups were included when at least 298 cases were available for IRT analyses. We were not able to include participants who self-described as Hispanic because the total sample never approached this number (n = 142

across all semesters). On the BDI-II, the number of participants who self-described as Asian American/Pacific Islander (subsequently referred to as Asian American in this article) also fell well below 300 (n = 241).

Total participants in each group for each measure were as follows, with the first number representing all analyses other than the IRT analyses (including factor analyses), and the second number representing the IRT analyses: (a) BDI-II, collected fall 2001 through spring 2004: African American participants (n = 400; 418), White participants (n = 1,051; 1,092); (b) PSWQ, collected spring 1999 through fall 2004, not collected spring 2003: African American participants (n = 814; 879), Asian American participants (n = 280; 298), White participants (n = 1,567; 1,657); (c) SIAS, collected fall 2000 through spring 2004: African American participants (n = 885; 967), Asian American participants (n = 386; 411), White participants (n = 2,332; 2,475). These samples are not independent.

In addition to their ethnicity, participants in all samples recorded their parents' years of education (up to 6 years of graduate school). Participants from fall 2000 and later recorded their household income on a 4-point scale: under $10,000, $11,000-$30,000, $31,000-$50,000, and over $50,000. Given that all participants were attending a university, it appeared useful to investigate how economic and educational data from our sample compare with national data. In all samples, the educational attainment of participants' parents was available. In the majority of the samples, household income was reported.

Because samples overlapped, we investigated these differences in the largest sample (for the SIAS) and then verified that effects were similar in the other samples; effects were substantively identical in the other samples. In the SIAS sample, income level category varied among individuals from different ethnic categories, as indicated by a significant chi-square value, $c2(6; N = 3,526) = 321.53$, $p < .001$; both Asian American participants and African American participants reported lower household income than White participants. For example, 55% of White participants reported a household income greater than $50,000, whereas only 32% of Asian American and 29% of African American participants reported that level of household income. For both African American and Asian American participants, the modal response was $11,000 to $30,000, but the percentages of participants reporting the higher two income categories was similar to the percentage giving the modal response (for African American participants: 32%, 29%, and 29%; for Asian American participants, 34%, 24%, and 32%). Asian American and African American participants did not show different income distributions, $c2(3; N = 1251) = 3.64$, $p = .30$. These data appear somewhat in contrast to national data from 2005; White participants in our sample reported a likely higher household income than the national median income in that ethnic group (about $51,000), whereas African American participants reported a somewhat higher level and Asian American participants a much lower level than the national median income (about $31,000 and about $61,000, respectively; DeNavas-Walt, Proctor, & Lee, 2006). Thus, in terms of household income and in comparison to their own ethnic groups on a national basis, our White participants are clearly advantaged, whereas our

African American participants appear somewhat advantaged and our Asian American participants appear somewhat disadvantaged.

Similarly, on average, African American participants reported that fathers' level of education was significantly less than that of White participants, $t(3155) = 7.05$, $p < .001$; both fathers' and mothers' years of education were fewer for Asian American participants than White participants ($ps < .01$), and mothers' years of education were fewer for Asian American participants than African American participants, $t(554) = 4.97$, $p < .001$. However, for all groups, mean level of parents' education was in the range of some college (e.g., range of means 15.93 [African American fathers] to 17.34 [White fathers]). Again, in comparison with national data from 2004, these data suggest that our White and African American participants may be advantaged compared to their own groups on the national level, whereas our Asian American participants may be somewhat disadvantaged. Only 17.4% of Africans American census respondents and 29.7% of White respondents 25 years or older report having a bachelor's degree or more, although 79.4% and 88.6% report high school or higher (U.S. Census Bureau, 2007b). These numbers suggest that, if our sample were representative, mean levels of education should be somewhat lower (e.g., closer to 12 years), but disparities between African American participants and White participants should be similar to those seen here. In contrast, nearly half of Asian American census respondents (48.4%) report a bachelor's degree or more (U.S. Census Bureau, 2007a), suggesting that our Asian American sample is somewhat disadvantaged in comparison to Asian Americans on a national level.

Measures

Beck Depression Inventory–Second Edition (BDI-II). The BDI-II (Beck, Steer, & Brown, 1996) is a 21-item self-report measure in which respondents are asked to rate the severity of cognitive, physiological, and affective symptoms of dysphoria they experienced over the previous 2 weeks on a scale ranging from 0 to 3. Items assess symptoms such as feelings of hopelessness, weight gain or loss, increases or decreases in amount of sleep, and suicidal ideation. Anchors for the scale vary for each item, although there are no reverse-scored items, and the scale always progresses from low to high severity. Scores on individual items are then summed to yield a total score. In the current sample, the BDI-II demonstrated good internal consistency when examined collectively or separately by ethnic group (all Cronbach's as > .89).

Social Interaction Anxiety Scale (SIAS). The SIAS (Mattick & Clarke, 1998) is a 20-item measure that employs a 0 (Not at All) to 4 (Extremely) Likert-type scale. The items describe anxiety-related reactions to a variety of social interaction situations, such as "I get nervous if I have to speak to someone in authority" and "When mixing socially I am uncomfortable." To score the SIAS, responses on each item are summed to yield a total score, which can range from 0 to 80. Three items are reverse-scored. Overall, research on the scale suggests good to excellent reliability and good construct and convergent validity (see Heimberg & Turk, 2002, for a review). The three reverse-scored items of the scale fail to load on the same factor as the other

items and must be accounted for by a separate factor or method factors (Rodebaugh, Woods, Heimberg, Liebowitz, & Schneier, 2006) and appear less related to social anxiety and more related to extraversion than is desirable (Rodebaugh, Woods, & Heimberg, 2007). Therefore, only the straightforwardly worded items were included in the IRT analyses. In the current sample, the straightforwardly worded items demonstrated good internal consistency when examined collectively or separately by ethnic group (all Cronbach's as > .91).

Penn State Worry Questionnaire (PSWQ). The PSWQ (Meyer et al., 1990) assesses the individual's level of pathological worry. Its 16 items (e.g., "I am always worrying about something" or "Once I start worrying, I cannot stop") are scored on a 1 (not at all typical of me) to 5 (very typical of me) Likert-type scale. Responses (including five reverse-scored items) are then summed to yield a total score, which can range from 16 to 80. The PSWQ has good internal consistency in both undergraduate (Meyer et al., 1990) and clinical (Brown, Antony, & Barlow, 1992) samples. Additionally, the PSWQ has demonstrated good test-retest reliability over periods up to 10 weeks (Meyer et al., 1990). Results regarding the factor structure of the PSWQ have been somewhat inconsistent, although many of the investigations appear to confirm that the factor structure is not straightforwardly unitary, because of differences in item wording, leading either to separate substantive factors, method factors, or some other method for accounting for item wording (Fresco et al., 2002; Hazlett-Stevens et al., 2004). Accordingly, we expected to examine only the straightforwardly worded items for DIF. In the current sample, the straightforwardly worded items demonstrated good internal consistency when examined collectively or separately by ethnic group (all Cronbach's as > .90).

Data-Analytic Specifications

Missing data. Analyses were conducted within each scale; therefore, decisions to retain or drop cases were based on missing data at the scale level. Between 4% and 10% of cases had partially missing data for a particular scale. A participant's data were retained for the factor analyses only if complete data were provided for the relevant scale, because the estimator (see below) for the factor analyses requires complete data. It should be noted here that the primary purpose of the study was to examine differences between the African American, Asian American, and White samples using DIF analyses, not to look at the factor structure of the questionnaires within the samples. In studies in which factor analyses are more central, multiple approaches to missing data, including multiple imputation (Rubin, 1987), can be used.

We retained a participant's data for the DIF analyses if the data contained no more than two missing responses on the relevant scale if that scale had seven or more items; for six or fewer items, we retained cases when there was no more than one missing response. We used these cutoffs in an attempt to ensure that data were only estimated for those participants who missed items at random.

Factor analyses. The sole purpose of the confirmatory factor analyses (CFAs) was to select a group of items from each scale that could be treated as unidimensional. The IRT analyses to follow require that the items analyzed be unidimensional, yet none of the scales we examined are actually unidimensional, despite the fact that they are often treated as if they are. The CFAs were conducted using the available literature as a guide. We excluded from consideration any models used to test method factors or correlated error variances. By method factor we mean a factor that consists only of items that have another loading on a substantive factor. For example, as described in the Measures section, some models for the SIAS have included a substantive social interaction anxiety factor, as well as method factors for item wording. In such cases, all variables loaded on the substantive factor and only reverse-scored questions cross-loaded onto a method factor to account for reverse-scored item wording. We did not consider such models because they could not be translated to the IRT analyses. When no clear factor solution was available, exploratory factor analyses (EFAs) were conducted in a randomly selected half of the total sample, followed by a CFA in the remaining half and in each ethnic subsample.

All factor-analytic methods we used are appropriate for ordinal data and implemented in the Mplus program (Muthén & Muthén, 1998-2004). All of the measures analyzed contain items with Likert-type response scales that are best considered ordered-categorical rather than continuous variables. EFA and CFA models were fitted to polychoric correlations using the robust weighted least squares estimator, WLSMV. WLSMV refers to a procedure by which parameters are estimated by weighted least squares using a diagonal weight matrix, and standard errors and the chi-square statistic are mean- and variance-adjusted using the full weight matrix.

For EFAs, emphasis was placed on the root mean square error of approximation (RMSEA, Steiger & Lind, 1980); a scree plot also was examined. For CFAs, global model fit was evaluated using the (a) Tucker–Lewis incremental fit index (TLI; Tucker & Lewis, 1973); (b) comparative fit index (CFI; Bentler, 1990); (c) root mean square residual (RMR); and the (d) standardized root mean square residual (SRMR; Bentler, 1995; Jöreskog & Sörbom, 1981). The magnitudes of these indices were evaluated with the aid of recommendations by Hu and Bentler (1999), Muthén and Muthén (1998-2004), and Yu and Muthén (as cited in Muthén & Muthén, 1998-2004). Because the primary purpose of these analyses was to isolate factors for use in the IRT analyses, we required that CFA models show at least reasonable fit.

Differential item functioning (DIF) analyses. IRT-based DIF testing was implemented to reveal both item discrimination and item threshold invariance across groups. The computer program IRT-LR-DIF (Thissen, 2001) was used to implement likelihood ratio–based DIF (LR-DIF, Thissen, Steinberg, & Gerrard, 1986), so-named because nested IRT models, differing with respect to their constraints, are compared using a likelihood ratio difference test. Samejima's (1969, 1997) graded model was fitted for responses to sets of unidimensional items, as identified in the factor analyses. The graded IRT model is a multivariate logistic regression model with one latent predictor (see Embretson & Reise, 2000, or Thissen & Wainer, 2001, for more about the graded model). One slope (discrimination) parameter and the appropriate number (response

options minus 1) of threshold parameters were estimated for each item; for example, for the SIAS items, which have five response options, one discrimination parameter and four threshold parameters were estimated.

Some of the items are treated as anchors to set a common scale for the latent variable between groups. Anchor items are assumed to be DIF-free; thus, they are not tested for DIF (and are not presented in our results). To identify a set of anchor items for each measure, we used a purification procedure suggested by Kim and Cohen (1995). The procedure begins with an omnibus test of DIF (testing DIF in a and b simultaneously) for each item. One item is tested at a time, with all other items treated as anchors. If no items show significant DIF, there is no need to proceed. Otherwise, the item with the largest statistically significant chi-square test statistic is eliminated, and the analysis is repeated with the remaining items. DIF testing is repeated as before, and again, the item with the largest statistically significant test statistic is eliminated. Items are eliminated one by one through repetition of this process until no further DIF is found. The final set of items showing no DIF serve as the anchors for the main analysis.

To test each (nonanchor) item for DIF, one fitting of the graded model with all item parameters for the anchors and the item under study constrained to be equal between ethnic groups was compared to a fitting with the parameters of the studied item free to vary between groups. If the test statistic for this omnibus comparison was significant ($p < .05$), group-invariance also was tested for the slope and thresholds separately. Thus, DIF could be detected for each item with respect to slope, thresholds, or both. It should be noted that a multigroup factor analysis using methods suited to categorical data could provide similar information, depending on the software used to conduct the analyses. We chose an IRT framework out of preference and the fact that the analyses were easier to conduct given the software we had available.

Results

Means

Means and standard deviations for each measure for the total sample and each ethnic group are presented in Table 1. Missing data were deleted listwise when constructing this table. As can be seen in the table, statistically significant differences in means were found for each scale across ethnic groups. However, these differences were generally small to moderate in terms of effect size; the largest effects (which were not large) were attributable to Asian American participants scoring higher on the SIAS and PSWQ than African American participants.

Confirmatory Factor Analyses

The purpose of these analyses was to select items that could be treated unidimensionally, a necessary step because the IRT analyses assume unidimensionality. Previous factor analyses for both the PSWQ (Hazlett-Stevens et al., 2004) and the SIAS (Rodebaugh et al., 2006) have indicated that the reverse-scored items of the measures must be accounted for to obtain

uniformly good fit. Although these previous studies focused primarily on method factors, such factors are not suitable for our subsequent analyses; we therefore focused on two-factor models in which the reverse-scored questions loaded on one factor that was permitted to correlate with a factor indicated by the straightforwardly worded items. Fit across the different sample groups is reported in Table 2. Two-factor fit for both the PSWQ and SIAS was acceptable to excellent across all fit indices. In the case of the PSWQ, it seems reasonable to assume that the difference between the factors is due to method variance rather than substantive differences (Brown, 2003), whereas in the case of the SIAS, both method variance and substantive variance has been found to vary given item wording (Rodebaugh et al., 2007). In both cases, we focus on the straightforwardly worded items in our IRT analyses because these items are less likely to be affected by error caused by participant confusion or carelessness (e.g., Spector, Van Katwyk, Brannick, & Chen, 1997) and, in the specific case of the SIAS, the reverse-scored questions appear to have undesirable relationships with other measures (Rodebaugh et al., 2007). Thus, reverse-scored items from both scales are not analyzed further.

Tables 1 & 2 are omitted from this formatted document.

The BDI-II represents a more challenging factoranalytic problem for several reasons. Virtually all previously published structural analyses, with the exception of Ward (2006), use methods that are more appropriate for continuous data than the ordered categories that make up the BDI items. In that literature, we could not find a consistent, replicable factor structure for the BDI-II (see Ward, 2006, for a relevant review). The factor analytic work of Ward also suggests that orthogonal decomposition of a general factor is the best-fitting factor model for the BDI-II; such a model is untenable for use in the IRT analyses we intended. Finally, although some researchers have chosen to treat the versions of the BDI as unidimensional for the purpose of DIF analyses (Bedi, Maraun, & Chrisjohn, 2001; Kim, Pilkonis, Frank, Thase, & Reynolds, 2002), we were unwilling to make this assumption without first testing our own data set.

For all of these reasons, we conducted EFAs on a randomly chosen half of each sample separately for the BDI-II. Our intention was to discover a reasonable factor structure that did not incorporate method factors or correlated errors terms, deleting items from the analyses if necessary. The factor structure of the EFAs was then confirmed separately with CFAs on the remaining half of each sample. Information from the EFAs for the BDI-II is reported in Table 3. The BDI-II analyses resulted in a three-factor solution, but this solution resulted in one factor indicated by only three items. Further, these three items all contained the phrase “loss of,” which suggests that the factor is likely to represent primarily method variance. We therefore tested a three-factor solution including the loss factor, but planned to drop these items for IRT analyses. Table 2 displays fit statistics for the BDI-II CFAs; the model fit reasonably to excellently well.

Table 3 is omitted from this formatted document.

We do not intend these factor solutions to be taken as strong indicators of the underlying structure of the BDI-II. Rather, our intention was to avoid the (in our view, incorrect) assumption of unidimensionality that would be inherent in submitting all items to the DIF analyses together. Nonetheless, some comment on the items associated with the factors may be useful. Item 11, and Items 16 through 20 loaded on the somatic factor. The other factor appeared to focus more on cognitive aspects of depression; Items 1 through 10 and Items 13 and 14 loaded on this factor. Both factors contained items that reflected affective aspects of depression. As noted by Ward (2006) in existing factor analyses of the BDI-II, so-called cognitive or somatic factors also contain affective or emotional content.

Differential Item Functioning

Between Ethnic Groups

DIF analyses. The DIF results are presented in Tables 4 to 8. All items for which some form of DIF (a and/or b) is given in boldface produced a statistically significant omnibus c2 ($p < .05$ following a Benjamini–Hochberg correction, Benjamini & Hochberg, 1995; Thissen, Steinberg, & Kuang, 2002). For each item, the discrimination (a) or threshold (b) parameters that are in boldface indicate which parameters drive the DIF. The items not displayed are anchor items and are assumed to contain no DIF. Note that a few items included in the tables do not contain statistically significant DIF. These items did not meet the standards for anchor items and also did not meet statistical significance after we applied the Benjamini–Hochberg correction. Thus, although they preliminarily indicated a tendency toward DIF, they ultimately did not meet more stringent criteria.

Table 4 displays DIF for the BDI-II items. Two BDI-II items showed a tendency to be more readily endorsed by African American participants when compared with White participants with the same degree of latent pathology. For example, African American participants had a greater likelihood of endorsing the more pathological response options for the punishment feelings item when compared to equally pathological White participants. Overall, very little DIF was found for the items, and, notably, no DIF was found for items identified as loading on the somatic factor.

Table 5 displays DIF for the PSWQ comparisons of White and African American participants. Seven of the 11 items contained DIF, 5 contained both discrimination and threshold DIF, 1 contained just discrimination DIF, and another contained just threshold DIF. The strongest effect here appeared to be for the 6 items that contained discrimination DIF. All discrimination DIF was negative, indicating that these items were less related to worry for African American participants than they were for White participants.

Table 6 displays DIF for PSWQ comparisons of White and Asian American participants. Threshold DIF was found for only two items, mostly working in opposite directions. Perhaps most notably, no discrimination DIF was found, indicating that these items measure worry similarly across these ethnic groups. In regard to lifelong worrying, positive threshold DIF was

found such that Asian American participants had a greater likelihood of endorsing more pathological response options.

Results for the SIAS are displayed in Tables 7 and 8. In comparisons of White participants with both African American participants and Asian American participants, most items displayed significant DIF. The results for the comparison with African American participants suggest that the items on the SIAS act differently across groups and, overall, less effectively for African American participants. The items do not distinguish well among African American participants with varying levels of latent social interaction anxiety. Three items showed discrimination DIF, and for each of these, the item was less effective in distinguishing among African American participants than it was among White participants. Furthermore, most items showed threshold DIF. For some items, this DIF was consistent. For example, for the item "I feel tense if I am alone with just one person," given any level of social interaction anxiety, African American participants had a greater likelihood of endorsing more pathological response options. No items showed the opposite pattern (African American participants having a lower likelihood of endorsing more pathological response options).

Table 4 is omitted from this formatted document.

Table 8 displays the comparable statistics for the comparison of White participants with Asian American participants. Overall, these results are more consistent. First, there was little evidence of discrimination DIF, except for Item 10, which concerns difficulty talking with other people; this item had stronger discriminating power for Asian American students. Although other discrimination DIF comparisons were not statistically significant, it is worth noting that all differences were in the same direction, indicating greater discrimination for Asian American participants than White participants. Threshold DIF was similarly consistent: With few exceptions, Asian American participants had greater a likelihood of endorsing more pathological responses at each level of social interaction anxiety. The only item not following this trend, for which Asian American participants had a greater likelihood of endorsing less pathological responses at each level of social interaction anxiety, was Item 20, concerning being unsure whether to greet someone. Also of note was the fact that only one of the items involving "mixing," all of which showed statistically significant or near statistically significant DIF for African American participants, showed any evidence of statistically significant DIF here.

Discussion

Differential item functioning may render scores on otherwise psychometrically strong instruments difficult or impossible to interpret meaningfully. Although African Americans in our study responded similarly to White individuals on the BDI-II, the measures of anxiety we examined (the SIAS and PSWQ) performed differently in both the African American and Asian American groups. As a consequence, using these anxiety measures for comparisons of mean differences in symptom levels or examination of factor structures between African American,

Asian American, and White samples may be problematic (albeit, to varying degrees). Further, comparisons between groups with Asian American and African American participants may be problematic when the relative make-up of the groups differs across samples. Thus, the measures may perform differently in a group of participants in which 10% are African American versus a group in which 30% are African American. The DIF found in this study varied by item, scale, and the particular ethnicity contrast conducted.

For the BDI-II, White and African American responses exhibited relatively little difference in the relationship between their responses and the relationship to the latent variable (e.g., discrimination DIF), differing on only two items. The BDI-II appears to measure dysphoria roughly equivalently across African American and White respondents. It is possible that with elimination of the two items demonstrating significant DIF (Punishment Feelings and Indecisiveness), the cross-ethnic applicability of this measure may be marginally improved.

The PSWQ performed less well in comparisons of White and Asian American students. Response patterns differed significantly on 4 of the 11 items that the study examined. Results for the African American versus White sample comparison were even more troublesome, suggesting significant threshold-related DIF for 7 of the 11 items. However, given that this DIF was primarily problematic at the lower end of the latent construct, it remains possible that the PSWQ would function as a relatively DIF-free instrument among clinical samples, in which higher levels of worry are expected. Further research would be necessary to confirm this possibility. At this point, the safest conclusion might be that the metric equivalence of the PSWQ permits its usage in clinical samples with participants from diverse ethnic backgrounds. The common use of this measure in nonclinical samples (e.g., undergraduates) may be more problematic, however, as the PSWQ might fail to detect subtle differences in lower levels of pathological worry patterns among individuals from different ethnic backgrounds (particularly among African American individuals).

Results for the SIAS suggest the response patterns of African American participants significantly differed from White participants. It is possible that these differences render the SIAS less effective in samples including African American participants; it seems more immediately likely that the differences would render accurate comparisons between ethnic groups difficult, if not impossible. The differences also did not appear to be consistent enough to allow for statistical correction or removal of particularly biased items.

Tables 5-8 are omitted from this formatted document.

Comparing White participants with Asian American participants on the SIAS suggested a different set of problems for the scale. Individual questions on the SIAS performed at least as well, if not somewhat better, in distinguishing Asian American individuals on the basis of social l interaction anxiety. The SIAS therefore appears to be just as precise (if not more so) an instrument among Asian Americans as White individuals. However, on the whole, Asian

American students were more likely to readily endorse pathological responses due to differential item properties. Thus, in our sample, the small mean differences between ethnic groups may be at least partially a result of measurement issues rather than a meaningful difference in the underlying construct of social interaction anxiety. This finding may also have consequences when the SIAS is used to select analogue groups of participants based on a cutoff score: the SIAS may have a tendency to select too many participants who belong to the Asian American ethnic group, because the cutoff is largely based on how the scale functions for White participants. Taken all together, the results of this study suggest that the SIAS should be used with caution in multiethnic undergraduate samples. Further study is necessary to determine the extent of the problems caused by the differential functioning we identified, as well as whether these findings extend to community or clinical samples. We do not claim to refute recent studies finding differences in level of social anxiety between Asian American and White college students (Horng & Coles, 2004; Lee et al., 2006; Okazaki, 2000; Okazaki et al., 2002). However, our results do open the possibility that some or all of the differences reported are due to biased items; the factors that biased the items of the SIAS might very well bias similar instruments. Only further research can answer this question definitively.

Our findings should be interpreted within the limitations of our samples and method. Because we used undergraduate samples, the application of these findings to clinical groups will be uncertain until further research is conducted. However, it is also worth noting that much of the extant reliability and validity data available on these questionnaires also comes from undergraduate samples. We note that our data on educational attainment of parents and household income suggests that both our White and African American groups are somewhat advantaged, which is unsurprising for undergraduates; however, economic and educational disparities persisted between these groups in a way that somewhat mirrors such disparities on a national level. Our Asian American sample is actually less advantaged than Asian American individuals on a national level. Such a finding is especially troublesome, given that some studies have suggested that income level is a better predictor of clinical anxiety level than ethnic background (Gillis et al., 1995; Kessler et al., 2005). We would have preferred to have had more participants, particularly in the Asian American subsample. Although we are aware of no standard method of determining power available for the analyses we present here, we suggest that it is safer to assume that the effects we have found, particularly regarding the Asian American sample, are relatively large. There may well be other effects that we did not have the power to detect. Our results for the BDI-II must be interpreted within the factor structure employed here. We believe these results are to be preferred over those obtained by treating the scales as unidimensional; at the same time, the factor structure we identified is not completely consistent with those in the extant literature—due in part, of course, to the fact that the factor structures in the extant literature are inconsistent.

The use of self-identification within broad categories to assess ethnic background also presents problems. We do not mean to imply that the broad ethnic definitions used here are ideal or that

research in this area should be confined to such groups. Although this approach is the most commonly used in the literature, self-identification within the broad categories listed here uses participant understandings of race as a stand-in for the more culturally complex constructs of ethnicity and identity, which can include family background, country of origin, heritage, years in the United States, acculturation, and the degree to which the individual identifies with these definitions (to name only a few such factors). The fact that statistically significant differences were found is all the more impressive given the acknowledged heterogeneity of the groups, which should tend to make finding significant DIF more difficult.

Despite these limitations, the results provide some clear indications regarding the relative status of these self-report measures in terms of invariance across ethnic groups. Prior to this study, there have been few large-scale investigations of these measures in African American and Asian American samples. IRT analyses allowed us to make specific comparisons between two groups and were able to identify specific trends in response decision making that can bias higher-order analyses, such as mean differences, single-group factor analysis, and regression analysis. The distinctions between these approaches can clearly be seen in the way that single-group factor analyses of the PSWQ failed to detect many of the areas of difference identified by the IRT analyses in this study.

Although we have focused on only three measures, we conjecture that the advantages and disadvantages of these measures are more the rule than the exception. As previously noted, participants from non-White ethnic groups are woefully underrepresented in clinical research, and there are not sufficient data on how individuals from different ethnic groups respond to these scales. When possible, we have suggested ways in which these measures can be improved for use in multiethnic samples, including the need for further study of the reliability and validity of the PSWQ and SIAS in community and clinical samples rather than the immediate discontinuation of use of these measures. We also believe that the key importance of this study is not the identification of flaws in a few measures, but instead the reinforcement of the need for careful examination of how response patterns might differ, not only as a function of racial category or ethnicity but also as a function of other areas of individual difference, such as gender or socioeconomic status. The extent of caution to be applied when using the PSWQ and the SIAS should be based on the pragmatic consequences of the DIF found in this study; further study in additional samples is necessary on this point.

We believe the clinical implications of this study are also important. The BDI-II, and to a lesser degree the PSWQ and SIAS, are frequently used to screen for the presence of symptoms in a variety of settings. Underestimation of scores can result in failure to detect individuals in need of services. Overestimation of scores can result in valuable resources being earmarked for populations in which there is less need.

We hesitate to draw large conceptual conclusions regarding these results in part because this is one of the first large-scale studies examining these measures. What the results do highlight is the

limitation of a purely etic approach to cross-cultural research. These measures have been treated as culturally equivalent because of researcher assumptions that the language and constructs are culture-neutral, and statistical findings that the groups do not significantly differ at a mean level. Our study clearly indicates that these questionnaires may not be as culturally neutral as was assumed, but we are limited in assessing what the potential sources of difference might be. Ideally, our questionnaire battery would have included emic measures designed to gauge the cultural background of the various subsamples, permitting a more fine-tuned analysis of the implications of these results.

The defining characteristic of an instrument is its appropriateness for the task for which it was designed. Often, these decisions are not binary, but contextual, based not only on what is the ideal tool for the task but also on what is the best tool available. Using common approaches of assessment of reliability and validity, the BDI-II, the SIAS, and the PSWQ have proven to be extremely effective tools in the measurement of depression, social anxiety, and worry. The goal of this study was to examine whether these measures were, in fact, adequate tools for the task of assessing symptoms across individuals of distinct ethnic backgrounds. As this study clearly demonstrates, these previously presumed psychometrically sound instruments have limitations in applicability that conventional approaches to assessment of reliability and validity might not detect. Importantly, future studies will be necessary to determine the impact the DIF found here has on the scales as a whole.3 In identifying some of their limitations, we hope both to suggest improvements in the times and ways in which these measures are used, as well as raise questions about ways in which these and similar instruments can be improved. The details can be crucial to refining the strategies we use to assess the needs of individuals from diverse backgrounds and refining the strategies of research and clinical practice. The goal, in the end, is to have tools that assess anxiety and depression equally well across people of all backgrounds.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

Notes

1. One point of confusion for the reader of cross-cultural research is the wide variety of labels applied to what, pragmatically, is the same group of people. We found an especially wide variance in the way studies described individuals of White/Caucasian/European American descent. For the sake of clarity, we have decided to use White as the main descriptor in this study. However, we do this with full awareness that it may be a faulty assumption to think that

individuals of these various ethnic descents respond equivalently to each of these descriptive terms and that the data from these studies are fully comparable.

2. Another problematic issue in the scientific literature is the distinction between race and ethnicity. Racial categories in the United States are used to refer to descendents of African, Asian, White, and Native American people; however, the use of racial categories has confounded a variety of factors associated with race, including socioeconomic status and acculturation. In keeping with recent recommendations, the term ethnicity or ethnic background has been offered as an attempt to better capture sociocultural factors, such as values, traditions, cultural practices, and language, which may better explain differences among racial and ethnic groups (Helms, Jernigan, & Mascher, 2005). In this article, we apply the terms ethnicity and ethnic background to connote that differences across racial categories may exist because of a wide variety of factors. This study involved cross-ethnic comparisons between African American, Asian American, and White American groups, and we recognize the diversity of ethnic practices within each category and the potential limitations associated with such a data-analytic approach (e.g. Japanese, Chinese descendents grouped into Asian American category; American-born and African-born individuals in the African American category).

3. We are examining this question in preparation for further publications. Our preliminary findings suggest that although there are differences between ethnic groups on the latent constructs, differential item functioning, in some cases, serves to distort and magnify these differences.

References

Alegria, M., Vila, D., Woo, M., Canino, G., Takeuchi, D., Vera, M., et al. (2004). Cultural relevance and equivalence in the NLAAS instrument: Integrating etic and emic in the development of cross-cultural measures for a psychiatric epidemiology and services study of Latinos. International Journal of Methods in Psychiatric Research, 13, 270-288.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories–IA and –II in psychiatric outpatients. Journal of Personality Assessment, 67, 588-597.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory (2nd ed.). San Antonio, TX: Psychological Corporation.

Bedi, R. P., Maraun, M. D., & Chrisjohn, R. D. (2001). A multisample item response theory analysis of the Beck Depression Inventory-1A. Canadian Journal of Behavioural Science, 33, 176-185.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57, 289-300.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238-246.

Bentler, P. M. (1995). EQS: Structural equations program manual, Version 5.0. Los Angeles: BMDP Statistical Software.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). Crosscultural research methods. New York: Wiley.

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? Behaviour Research and Therapy, 41, 1411-1426.

Brown, T. A., Antony, M. M., & Barlow, D. H. (1992). Psychometric properties of the Penn State Worry Questionnaire in a clinical anxiety disorders sample. Behaviour Research and Therapy, 30, 33-37.

Carmody, D. P. (2005). Psychometric characteristics of the Beck Depression Inventory-II with college students of diverse ethnicity. International Journal of Psychiatry in Clinical Practice, 9, 22-28.

Carter, M. M., Sbrocco, T., Miller, O., Jr., Suchday, S., Lewis, E. L., & Freedman, R. E. K. (2005). Factor structure, reliability, and validity of the Penn State Worry Questionnaire: Differences between African-American and White-American college students. Journal of Anxiety Disorders, 19, 827-843.

Dutton, G. R., Grothe, K. B., & Jones, G. N., Whitehead, D., Kendra, K., & Brantley, P. J. (2004). Use of the Beck Depression Inventory-II with African American primary care patients. General Hospital Psychiatry, 26, 437-442.

DeNavas-Walt, C., Proctor, B. D., & Lee, C. H. (2006). Income, poverty, and health insurance coverage in the United States: 2005. Washington, DC: Government Printing Office.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.

Fresco, D. M., Heimberg, R. G., Mennin, D. S., & Turk, C. L. (2002). Confirmatory factor analysis of the Penn State Worry Questionnaire. Behaviour Research and Therapy, 40, 313-323.

Gillis, M. M., Haaga, D. A. F., & Ford, G. T. (1995). Normative values for the Beck Anxiety Inventory, Fear Questionnaire, Penn State Worry Questionnaire, and Social Phobia and Anxiety Inventory. Psychological Assessment, 7, 450-455.

Grothe, K. B., Dutton, G. R., & Jones, G. N. (2005). Validation of the Beck Depression Inventory-II in a low-income African American sample of medical outpatients. Psychological Assessment, 17, 110-114.

Hazlett-Stevens, H., Ullman, J. B., & Craske, M. G. (2004). Factor structure of the Penn State Worry Questionnaire: Examination of a method factor. Assessment, 11, 361-370.

Heimberg, R. G., & Turk, C. L. (2002). Assessment of social phobia.

In R. G. Heimberg & E. Becker (Eds.), Cognitive-behavioral group therapy for social phobia: Basic mechanisms and clinical strategies (pp. 107-126). New York, NY: Guilford Press.

Helms, J. E., Jernigan, M., & Mascher, J. (2005). The meaning of race in psychology and how to change it: A methodological perspective. American Psychologist, 60, 27-36.

Hong, J. J., & Woody, S. R. (2007). Cultural mediators of selfreported social anxiety. Behaviour Research and Therapy, 45, 1779-1789.

Horng, B., & Coles, M. E. (2004, March). Investigating the link between social anxiety and maladaptive perfectionism in Asian Americans. Poster presented at the annual meeting of the Anxiety Disorders Association of America, Miami, FL.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.

Jöreskog, K. G., & Sörbom, D. (1981). LISREL V: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: National Educational Resources.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and ageof- onset distributions of DSM-IV disorders in the National Comorbidity Survey replication. Archives of General Psychiatry, 62, 593-602.

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chisquare,

Raju's area measures, and the likelihood ratio test on detection of differential item functioning. Applied Measurement in Education, 8, 291-312.

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. Psychology and Aging, 17, 379-391.

Lam, C. Y., Pepper, C. M., & Ryabchenko, K. A. (2004). Case identification of mood disorders in Asian American and Caucasian American college students. Psychiatric Quarterly, 75, 361-373.

Lee, M. R., Okazaki, S. & Yoo, H. C. (2006). Frequency and intensity of social anxiety in Asian Americans and European Americans. Cultural Diversity and Ethnic Minority Psychology, 12, 291-305.

Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. Behaviour Research and Therapy, 36, 455-470.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. Behaviour Research and Therapy, 28, 487-495.

Muthén, L. K., & Muthén, B. O. (1998-2004). Mplus user's guide (3rd ed.). Los Angeles: Muthén & Muthén. Neal, A., & Turner, S. (1991). Anxiety disorders research with African Americans: Current status. Psychological Bulletin, 109, 400-410.

Okazaki, S. (1997). Sources of ethnic differences between Asian American and White American college students on measures of depression and social anxiety. Journal of Abnormal Psychology, 106, 52-60.

Okazaki, S. (2000). Asian American–White American difference on affective distress symptoms: Do symptom reports differ across reporting methods? Journal of Cross-Cultural Psychology, 31, 603-625.

Okazaki, S., Liu, J. F., & Longworth, S. L. (2002). Asian American–White American differences in expressions of social anxiety: A replication and extension. Cultural Diversity and Ethnic Minority Psychology, 8, 234-247.

Rodebaugh, T. L., Woods, C. M., & Heimberg, R. G. (2007). The reverse of social anxiety is not always the opposite: The reverse-scored items of the Social Interaction Anxiety Scale do not belong. Behavior Therapy, 38, 192-206.

Rodebaugh, T. L., Woods, C. M., Heimberg, R. G., Liebowitz, M. R., & Schneier, F. R. (2006). The factor structure and screening utility of the Social Interaction Anxiety Scale. Psychological Assessment, 18, 231-237.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34(4 Pt. 2), 100.

Samejima, F. (1997). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. Psychometrika, 62, 471-493.

Scott, E. L., Eng, W., & Heimberg, R. G. (2002). Ethnic differences in worry in a nonclinical population. Depression and Anxiety, 15, 79-82.

Snowden, L. R. (2001). Barriers to effective mental health services for African Americans. Mental Health Services Research, 3, 181-187.

Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don‘t reflect two constructs: How item characteristics can produce artifactual factors. Journal of Management, 23, 659-677.

Steiger, J. H., & Lind, J. C. (1980, May). Statistically-based tests for the number of factors. Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood ratio tests for differential item functioning. University of North Carolina at Chapel Hill: L. L. Thurstone Psyhometric Laboratory.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond groupmean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. Journal of Educational and Behavioral Statistics, 27, 77-83.

Thissen, D., & Wainer, H. (Eds.). (2001). Test scoring. Mahwah, NJ: Lawrence Erlbaum.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

U.S. Census Bureau. (2007a). The American community: Asians: 2004. Retrieved May 16, 2007, from http://www.census.gov/prod/2007pubs/acs-05.pdf

U.S. Census Bureau. (2007b). The American community: Blacks: 2004. Retrieved May 16, 2007 from http://www.census.gov/prod/2007pubs/acs-04.pdf

Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory–II. Psychological Assessment, 18, 81-88.

Wells, K., Klap, R., Koike, A., & Sherbourne C. (2001). Ethnic disparities in unmet need for alcoholism, drug abuse, and mental health care. American Journal of Psychiatry, 158, 2027-2032.